音声認識用言語モデルのための タスク適応化と定型表現の利用

中川　聖一[†]　赤松　裕隆[†]　西崎　博光[†]

本研究では大規模コーパスが利用可能な新聞の読み上げ音声の認識のための精度の良い言語モデルの構築を実験的に検討した．N-gram言語モデルの改善を目指し，以下の3つの点に注目した．まずN-gram言語モデルはタスクに依存するので，タスクに関する大量のデータベースを用いて構築される必要があることに注目し，共通の大量データベースによる言語モデルをもとに，同一ジャンルの過去の記事を用いるタスク適応化の方法とその有効性を示す．次に，新聞記事は話題が経時的に変化するので，数日間～数週間の直前の記事内容で言語モデルの適応化を行なう方法とその有効性を示す．最後に新聞テキストには，使用頻度の高い(特殊)表現や，固定的な言い回しなどの表現(以下，定型表現と呼ぶ)が多いことに注目し，複数形態素から成る定型表現を抽出し，これを1形態素として捉えた上で，N-gram言語モデルを構築する方法を検討し，有用性を示す．

キーワード:　音声認識，言語モデル，*N-gram*，タスク適応化，定型表現

A task adaptation method and use of idiomatic expression of stochastic language model for speech recognition

SEIICHI NAKAGAWA[†] , HIROTAKA AKAMATSU[†]
and HIROMITSU NISHIZAKI[†]

In this paper, we describe a method that constructs language models using a task-adaptation strategy and idiomatic expressions of news articles. To build an effective N-gram based language model, it should be noted that the training data must be prepared as much as possible. However, for a given task/topic, it is very difficult to gather much data. First, we investigated the effect of a task adaptation method of N-gram language model using a limited amount of target articles. Second, we investigated the effect of the language model adaptation method using the latest articles. Third, we investigated the effect of the use of idiomatic expressions as morpheme units, since some specific expressions and idiomatic expressions are frequently observed in news articles. We show our proposed three methods are effective for constructing N-gram language models.

KeyWords:　*Speech Recognition, Language Model, N-gram, Task Adaptation, Idiomatic Expression*

† 豊橋技術科学大学工学部情報工学系, Dept. of Information and Computer Sciences, Faculty of Eng., Toyohashi University of Technology

1　はじめに

近年の著しい計算機速度の向上，及び，音声処理技術/自然言語処理技術の向上により，音声ディクテーションシステムやパソコンで動作する連続音声認識のフリーソフトウェアの公開など，音声認識技術が実用的なアプリケーションとして社会に受け入れられる可能性がでてきた(西村, 伊藤, 山崎, 萩野 1998; 甲斐, 伊藤, 山本, 中川 1997)．我が国では，大量のテキストデータベースや音声データベースの未整備のため欧米と比べてディクテーションシステムの研究は遅れていたが，最近になって新聞テキストデータやその読み上げ文のデータが整備され(伊藤 1997)，ようやく研究基盤が整った状況である．

このような背景を踏まえ，本研究では大規模コーパスが利用可能な新聞の読み上げ音声の精度の良い言語モデルの構築を実験的に検討した．音声認識のためのN-gram言語モデルでは，N=3〜4で十分であると考えられる(大附, 吉田, 松岡, 古井 1997; 森, 山地 1997a; Woodland *et al.* 1997)．しかし，N=3ではパラメータの数が多くなり，音声認識時の負荷が大きい．そこで，大語彙連続音声認識では，第1パス目はN=2のbigramモデルで複数候補の認識結果を出力し，N=3のtrigramで後処理を行なう方法が一般的である．本研究では，第2パスのtrigramの改善ばかりでなく，第1パス目のbigram言語モデルの改善を目指し，以下の3つの点に注目した．

まずタスクについて注目する．言語モデルをN-gramベースで構築する場合(ルールベースで記述するのとは異なり)，大量の学習データが必要となる．最近では各種データベースが幅広く構築され，言語モデルの作成に新聞記事などの大規模なデータベースを利用した研究が行なわれている (大附, 森, 松岡, 古井, 白井 1995)．しかし N-gram はタスクに依存するのでタスクに関する大量のデータベースを用いて構築される必要がある．例えば，観光案内対話タスクを想定し，既存の大量の言語データに特定タスクの言語データを少量混合することによって，N-gram言語モデルの性能の改善が行なわれている (伊藤, 牧野 1996)．また，複数のトピックに関する言語モデルの線形補間で適応化する方法が試みられている (Marlin, Liermann 1997)．~~本研究ではタスクへの適応化のために，同一ジャンルの過去の記事を用いる方法とその有効性を示す．~~

次に言語モデルの経時変化について注目する．例えば新聞記事などでは話題が経時的に変化し，新しい固有名詞が短期的に集中的に出現する場合が多い．以前の研究では、直前の数百単語による言語モデルの適応化(キャッシュ法)が試みられ(Kuhn, Mori 1990)，小さいタスクではその有効性が示されてはいるが，~~本論文では直前の数万〜数十万語に拡大する．つまり，直前の数日間〜数週間の記事内容で言語モデルを適応化する方法を検討し，その有効性を示す．~~

最後に認識単位に注目する．音声認識において，認識単位が短い場合認識誤りを生じやすく，付属語においてその影響は大きいと考えられ，小林らは，付属語列を新たな認識単位とした場合の効果の検証をしている(小林, 中野, 和田, 小林 1998)．また高木らは，高頻度の付属語連鎖，関連率の高い複合名詞などを新しい認識単位とし，これらを語彙に加えることによる言語モデルの性能に与える影響を検討している(小黒, 高木, 橋本, 尾関 1998)．なお，連続する単語クラス

を連結して一つの単語クラスとする方法や句を一つの単位とする方法は以前から試みられているが，いずれも適用されたデータベースの規模が小さい(Giachin 1995; 政瀧, 松永, 匂坂 1995)．同じような効果を狙った方法として，N-gramのNを可変にする方法も試みられている(Marlin, Liermann 1997)．なお，定型表現の抽出に関する研究は，テキスト処理分野では多くが試みられている(例えば，新納, 井佐原 1995; 北, 小倉, 森本, 矢野 1995)．

新聞テキストには，使用頻度の高い(特殊)表現や，固定的な言い回しなどの表現(以下，定型表現と呼ぶ)が非常に多いと思われる．定型表現は，音声認識用の言語モデルや音声認識結果の誤り訂正のための後処理に適用できる．そこでまず，定型表現を抽出した．次に，これらの(複数形態素から成る)定型表現を1形態素として捉えた上で，N-gram言語モデルを構築する方法を検討する．評価実験の結果，長さ2および3以下である定型表現を1形態素化してbigram, trigram言語モデルを作成することで，bigramに関しては，エントロピーが小さくなり，言語モデルとして有効であることを示す．

なお，これらの手法に関しては様々な方法が提案されているが，大規模のテキストデータを用いて，タスクの適応化と定型表現の導入の有効性を統一的に評価した研究は報告されていない．

2　言語モデルの評価基準

2.1　エントロピーとパープレキシティ

言語モデルの評価基準として，エントロピーとパープレキシティを用いる．エントロピーとパープレキシティは共に，対象とする文集合の複雑さを定量的に示す指標で，その文集合が複雑なほど，それぞれの値は大きくなる．

単語列を生成する情報源をモデル化したものを言語モデルと呼ぶ．いま言語Lにおいて，文(単語列)$W_i = w_1 \cdots w_{L_i}$の出現確率を$P(W_i)$とすれば，文集合$W_1, W_2, \cdots, W_N$のエントロピーは次式で求められる．

$$H(L) = -\sum_{i=1}^{N} P(W_i) \log P(W_i) \tag{1}$$

テキスト文の連接を$W = W_1 W_2 \cdots W_N = w_1 w_2 \cdots w_T$とすれば，テストセットのエントロピーは

$$H(L) = -\log P(W) \tag{2}$$

で示される．トライグラムを用いた場合，P(W)は

$$\begin{aligned} P(W) &= P(W_1)P(W_2)\cdots P(W_N) \\ &= P(w_1|*\ \#)P(w_2|\#\ w_1) \\ &\quad P(w_3|w_1\ w_2)\cdots P(w_T|w_{T-2}\ w_{T-1}) \end{aligned} \tag{3}$$

となる(注:#は文頭を，*は文末を示す．以降の評価実験では句読点を含む)．

この時，一単語当たりのエントロピーは

$$H_0(L) = -\frac{\sum_i \log P(W_i)}{\sum_i L_i} \tag{4}$$

また，言語の複雑さ・パープレキシティは

$$PP = 2^{H_0(L)} \tag{5}$$

と定義される．

パープレキシティは，情報理論的にある単語から後続可能な単語の種類数を表している．この値が大きくなるほど，単語を特定するのが難しくなり，言語として複雑であるといえる．また逆に，この値が小さくなるほど，音声認識での後続予測単語を特定するのがやさしくなるので，認識率が上がる傾向にある(中川 1992)．

日本語の単語の定義は定かでなく，また形態素の定義も異なる．そこで，本論文では文字単位のエントロピー(パープレキシティ)の指標も用いる．

2.2　補正パープレキシティ

本研究で使用したCMU SLM toolkit(Rosenfeld 1995) では語彙に含まれないものは全て一つの未知語のカテゴリにまとめられ，語彙に含まれる形態素と等価に未知語のカテゴリは扱われる．そのため語彙サイズのセットが小さい程(カバー率が小さい程)，パープレキシティは小さくなるということになり好ましくない．そこで評価テキスト中に出現した未知語の種類 m と，未知語の出現回数 n_u を用いてパープレキシティを補正する(Ueberla 1994)．補正パープレキシティは

$$APP = (P(w_1 ... w_n) m^{-n_u})^{-\frac{1}{n}} \tag{6}$$

で与えられる．これは，複数の未知語はそれぞれ等確率に生じると仮定して，補正したものである．勿論，これは評価テキストの大きさに依存する(テキストが大きくなると未知語の種類が増える)ので，簡易的な補正である．より厳密には未知語に対しては出現頻度を考慮するか(中川, 赤松 1998)，未知語の生成モデルを用いる必要がある(森 1997a)．なお，一般には，未知語部分はスキップしてパープレキシティを算出する方法がよく使われている．

3　言語モデルの適応化

3.1　面種別での学習と評価

タスク依存の言語モデルを構築する場合，ターゲットとするタスクに関するデータのみを用いて学習する方がよいと考えられる(松永, 山田, 鹿野 1991)．

学習と評価用のコーパスとして毎日新聞の1991年～1994年の記事を用いた．形態素解析にはRWCPが提供している毎日新聞形態素解析データを，電総研で作成された括弧除去ツールで

加工し，使用している(伊藤,松岡,竹沢,武田,鹿野 1996)．学習には1991年1月から1994年11月までの記事を用い，評価には1994年12月の記事を用いた．毎日新聞には全部で13面種に分類されているが,「社説」,「科学」,「読書」などの面種にはデータが少な過ぎるので，面種別の結果は省いた．登録した形態素数は5000 , 20000 の 2 通りで， bigram , trigram の学習と評価に CMU SLM toolkit を使用した．表1に用いたコーパスの諸量をまとめた(学習テキストが1994年1月～11月の場合の結果は，文献(赤松,中川 1997)を参照されたい)．

これらのデータを用いて作成したbigramとtrigramの評価結果を表2,3に示す．紙数の関係で，5000形態素に関する結果は省略した．これらの表では，実験結果をエントロピーではなくパープレキシティで表示している．これは音声認識実験を行なうことを踏まえ，情報理論的にある単語から後続可能な単語の種類数を示すパープキシティという指標の方が直観的にわかりやすいためである．

またカバー率とは，unigram のヒット率のことである．これらの結果より以下の事がわかる．

- bigramとtrigramを比較すると，bigramより，trigramで言語モデルを構築した方が，トレーニングデータとテストデータのどちらのパープレキシティも小さくなる．
- テストデータとトレーニングデータを比較すると，形態素数 5000 の bigram では，テストデータとトレーニングデータとの間にパープレキシティの差はほとんど見られなかった．しかし，それ以外の言語モデル(bigram 形態素数 20000, trigram 形態素数 5000, trigram 形態素数 20000) では，テストデータとトレーニングデータとの間にパープレキシティの差が大きい．これは補正パープレキシティでも同様である．特に形態素数 20000 の trigram で差が大きい．これは，9000万形態素では，トレーニングデータ量が不足していることを示している．
- 全面種で学習した場合と面種別で学習した場合の比較をすると，面種別に語彙を設定する方がカバー率は向上する．また，テストデータのパープレキシティに関しては，形態素数 20000 の bigram では全面種で学習するより面種別で学習する方がパープレキシティが小さくなる．trigramでは面種別で学習するより全面種で学習する方がパープレキシティが小さくなる．これは，面種別ではトレーニングデータが不足することによると考えられる．なお，テストデータの補正パープレキシティに関しては，形態素数 20000 の trigramでは面種別で学習するより全面種で学習する方が補正パープレキシティが小さくなる．bigramでは全面種で学習するより面種別で学習する方が補正パープレキシティが小さくなる．

  また，スポーツ面に関しては全面種で学習するより，面種別(すなわち，スポーツ面)で学習する方がパープレキシティが小さくなる傾向が見られる．これは，スポーツ面は他の面種と異なった文が多いことによる．

表には示さなかったが，形態素数 5000 の bigram に関しては，全面種での学習では 4 年分

の新聞記事で十分な学習が出来ている．一方，面種別での学習ではトレーニングデータとテストデータのパープレキシティの間に差があるのでトレーニングデータの不足が見られる．しかしトレーニングデータの不足が見られるものの，全面種で学習した言語モデルより面種別で学習した言語モデルの方がテストデータのパープレキシティが小さい．つまり，全面種で学習した言語モデルより面種別で学習した言語モデルを使用する方がよいことになる．形態素数 5000 の trigram に関しては，面種別学習による効果はパープレキシティでは見られないが，補正パープレキシティでは効果が見られる．

形態素数 20000 の trigram に関しては，トレーニングデータとテストデータの(補正)パープレキシティの比較によって，面種別での学習のみならず全面種での学習でもトレーニングデータ量の不足が起きていることが分かる．全面種で学習した言語モデルと面種別で学習した言語モデルをテストデータの(補正) パープレキシティで比較すると，形態素数 5000 の bigram での比較とは逆に，面種別で学習した言語モデルより全面種で学習した言語モデルの方が，(補正) パープレキシティが小さく，面種別で学習した言語モデルを使用するより，全面種で学習した言語モデルを使用する方がよいという結論が得られた．

以上から，形態素数 5000 の bigram を言語モデルに使用する場合は，面種別で学習した言語モデルを用いればよいことがわかった．しかし，最近の大語彙音声認識に用いられる形態素数は20000以上で，また第2パスに言語モデルとして trigram を使用するのが主流となりつつある．形態素数 20000 の trigram だと，本研究で用いたトレーニングデータ量程度では，面種別で学習した言語モデルを使用するより，全面種で学習した言語モデルを使用する方がよい．そこで，タスク(新聞では，面種)依存のより精度のよい言語モデルを構築するために全面種の記事で構築した言語モデルを，ターゲットとするタスク(面種)に適応化する手法をとる必要がある．

表 1　新聞記事のコーパス

面種	トレーニングデータ			テストデータ		
	文字数	形態素数	形態素種類数	文字数	形態素数	形態素種類数
解説	3976931	2476575	50091	101307	63528	8348
社説	4877489	3019306	46278	159906	100158	10129
国際	9585733	5782838	61140	207001	125675	10084
経済	10710679	6453508	71807	289227	173612	13142
特集	4131569	2532572	62765	163926	101799	12729
総合	20655857	12574265	132921	478066	291686	23800
家庭	6889896	4237176	72178	182299	113454	12041
文化	2333331	1434018	52125	81204	50018	8255
読書	1366333	846721	38937	32234	19908	4508
科学	769526	467918	20794	14700	9031	2068
芸能	3522025	2118354	60797	67238	40386	7139
スポーツ	8477958	5254979	57446	260096	160158	11642
社会	35538841	22470091	140027	701295	443885	24712
全面種	147654658	91273769	307557	3314722	2051584	56210

表 2　新聞記事の bigram(20000 形態素)

(a) トレーニングデータ (括弧内は文字単位のパープレキシティ)

面種	全面種で学習, 面種別で評価				面種別で学習, 面種別で評価			
	PP	APP	未知語の種類	カバー率 (%)	PP	APP	未知語の種類	カバー率 (%)
国際	86.0 (14.7)	113.1 (17.3)	42998	97.4	67.2 (12.7)	77.9 (13.8)	41140	98.6
経済	89.1 (15.0)	119.9 (17.9)	53015	97.3	68.7 (12.8)	81.6 (14.2)	51807	98.4
家庭	103.2 (17.3)	163.3 (23.0)	53389	95.8	80.0 (14.8)	106.3 (17.6)	52178	97.4
スポーツ	112.9 (18.7)	171.7 (24.3)	40204	96.0	70.8 (14.0)	81.9 (15.3)	37446	98.6
社会	86.8 (16.8)	129.0 (21.6)	120090	96.6	75.9 (15.4)	106.8 (19.2)	120027	97.1
全面種	91.0 (16.3)	136.7 (20.9)	287557	96.8				

(b) テストデータ (括弧内は文字単位のパープレキシティ)

面種	全面種で学習, 面種別で評価				面種別で学習, 面種別で評価			
	PP	APP	未知語の種類	カバー率 (%)	PP	APP	未知語の種類	カバー率 (%)
国際	92.8 (15.7)	112.9 (17.6)	2140	97.4	89.2 (15.3)	101.6 (16.5)	1658	98.2
経済	102.8 (16.1)	132.5 (18.8)	3751	96.9	100.0 (15.9)	119.5 (17.7)	2985	97.8
家庭	114.7 (19.1)	170.3 (24.5)	3852	95.2	113.2 (19.0)	148.6 (22.5)	3031	96.6
スポーツ	121.4 (19.2)	168.8 (23.5)	3602	96.0	98.2 (16.9)	112.4 (18.3)	2165	98.2
社会	96.3 (18.0)	132.4 (22.0)	9737	96.5	93.1 (17.6)	123.8 (21.1)	9258	96.9
全面種	105.6 (17.9)	155.8 (22.8)	36731	96.3				

表 3　新聞記事の trigram(20000 形態素)

(a) トレーニングデータ (括弧内は文字単位のパープレキシティ)

面種	全面種で学習, 面種別で評価				面種別で学習, 面種別で評価			
	PP	APP	未知語の種類	カバー率 (%)	PP	APP	未知語の種類	カバー率 (%)
国際	30.6 (7.9)	40.2 (9.3)	42998	97.4	20.9 (6.3)	24.2 (6.8)	41140	98.6
経済	30.1 (7.8)	40.6 (9.3)	53015	97.3	20.1 (6.1)	23.9 (6.8)	51807	98.4
家庭	36.0 (9.1)	57.0 (12.0)	53389	95.8	23.2 (6.9)	30.8 (8.2)	52178	97.4
スポーツ	31.9 (8.6)	48.6 (11.1)	40204	96.0	19.3 (6.3)	22.3 (6.9)	37446	98.6
社会	27.0 (8.0)	40.1 (10.3)	120090	96.6	20.9 (6.8)	29.3 (8.5)	120027	97.1
全面種	29.7 (8.1)	44.6 (10.5)	287557	96.8				

(b) テストデータ (括弧内は文字単位のパープレキシティ)

面種	全面種で学習, 面種別で評価				面種別で学習, 面種別で評価			
	PP	APP	未知語の種類	カバー率 (%)	PP	APP	未知語の種類	カバー率 (%)
国際	53.7 (11.2)	65.4 (12.7)	2140	97.4	62.7 (12.3)	71.4 (13.3)	1658	98.2
経済	59.2 (11.6)	76.4 (13.5)	3751	96.9	69.5 (12.8)	83.0 (14.2)	2985	97.8
家庭	72.8 (14.4)	108.0 (18.4)	3852	95.2	86.0 (16.0)	113.0 (19.0)	3031	96.6
スポーツ	67.4 (13.4)	93.7 (16.4)	3602	96.0	63.8 (12.9)	73.0 (14.0)	2165	98.2
社会	52.4 (12.3)	72.0 (15.0)	9737	96.5	56.7 (12.9)	75.4 (15.4)	9258	96.9
全面種	61.3 (12.8)	90.5 (16.3)	36731	96.3				

3.2 適応化法

新聞記事では数日間に渡って関連のある記事が載っていることがある．そこで記事の評価時に，過去の数日間の記事で言語モデルを適応化しておけば，適応前より精度のよい言語モデルが出来ると考えられる．

ここで，N-gram 言語モデルの適応化にはMAP推定 (最大事後確率推定)(伊藤 1996; Federico 1997; 政瀧, 匂坂, 久木, 河原 1997; 赤松 1997) を用いる．適応化サンプルを与えた後の推定値は次式で与えられ，推定前の条件確率と現在与えたサンプルとの間で，サンプル数で重み付けされた線形補間の形になっている．

$$prob = \frac{\alpha \cdot N_0 \cdot prob_0 + N_1 \cdot prob_1}{\alpha \cdot N_0 + N_1} \tag{7}$$

α	**重み**
N_0	**標準言語モデルの総数**
N_1	**適応化サンプルの総数**
$prob$	MAP **推定後の条件確率** (N-gram **確率**)
$prob_0$	**標準言語モデルでの条件確率**
$prob_1$	**適応化サンプルでの条件確率**

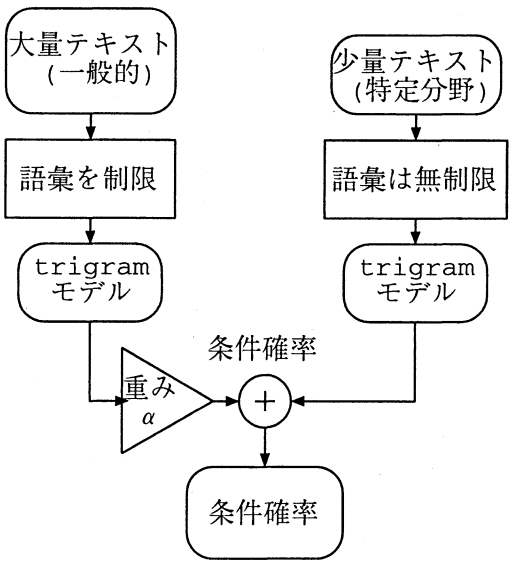

図 1　MAP推定のブロック図

今回の実験では標準言語モデルと適応化サンプルによる言語モデルの 2 つを構築しておき，バックオフを行なってスムージングした 2 つの条件確率を用いてMAP 推定を行なっている．こ

の過程のブロック図を図1に示す．標準言語モデルでは，新聞記事の全面種に対応する学習サンプルで出現頻度の高い形態素20000に限定した．適応化サンプルでは語彙を限定せず，全ての形態素を語彙リストに登録した．そのため，2つのモデルの語彙リストは独立している．

実験手順としては，①形態素数20000の標準言語モデル(trigram)を構築し，②標準言語モデルを事前モデルとして，面種別の適応化サンプルで ターゲットタスクの言語モデルをMAP推定し，③テストデータのパープレキシティを求める．本実験では，αを種々変えてパープレキシティが最小となる場合を求めた．

3.3 実験結果

実験結果を表4,5に示す．最適なαの値は5 日間の適応化データに対してはほとんどの面種で0.01，14日間の適応化データに対しては0.02~0.04であり，ほぼデータ量に比例した．これらの表より

- 適応化前より適応化後の方がパープレキシティが小さくなること
- 5日より14日間の適応化サンプルの方がパープレキシティが小さくなること
- 6カ月前の数日間より直前の数日間の記事での適応化の方がパープレキシティが小さくなること

が分かる．通常，直前の数百単語をキャッシュとして用いて適応化する方法が効果があると言われているが(Kuhn 1990)，これよりも大量の直前データを用いる方が効果があるということである(Marlin, Liermann 1997)．特に，スポーツ面において，直前の記事による適応化の効果が大きい．これは，他の面種記事よりも特定の話題が短期間継続するためと考えられる．

国際面とスポーツ面で適応化サンプルの期間を5,14日,1,2,3,6カ月にして求めたパープレキシティと補正パープレキシティを図2に示す．これを見ると，適応化サンプルの量が多くなるほど，パープレキシティが小さくなること，日数が多くなるにつれてパープレキシティが飽和していくことが分かる．

また，直前の適応化データと6カ月前の適応化データを比べると，後者の場合の方がやや最適なαの値が大きくなった．これは直前の適応化データの方が6カ月前の適応化データよりも有用であることを示している．

3.4 固有名詞の適応化

前述したように，新聞記事では数日間に渡って関連のある記事が載っていることが多い．音声認識では特に固有名詞の扱いが重要となってくるので，固有名詞の登録法について検討した．固有名詞はトピックに依存するものが多いので，数日間に渡って局所的に出現する傾向があると考えられる．そこで数日間~数週間中に出現した固有名詞を基本語彙に追加することにより，評価文の固有名詞をどの程度カバーすることが出来るかを調べた．

表 4　パープレキシティー(面種別，20000形態素)
(括弧内は文字単位のパープレキシティ)

記事・面	適応前	直前の記事で適応		6カ月前の記事で適応	
		5日分	14日分	5日分	14日分
国際	53.7 (11.2)	50.6 (10.8)	50.5 (10.8)	52.1 (11.0)	52.4 (11.1)
経済	59.3 (11.6)	56.3 (11.2)	55.9 (11.2)	57.6 (11.4)	57.7 (11.4)
家庭	72.8 (14.4)	67.7 (13.8)	68.0 (13.8)	68.8 (13.9)	69.6 (14.0)
スポーツ	67.4 (13.4)	59.5 (12.4)	57.1 (12.1)	63.3 (12.9)	62.9 (12.8)
社会	52.4 (12.3)	50.4 (12.0)	50.0 (11.9)	51.9 (12.2)	51.9 (12.2)

表 5　補正パープレキシティー(面種別，20000形態素)
(括弧内は文字単位の補正パープレキシティ)

記事・面	適応前	直前の記事で適応		6カ月前の記事で適応	
		5日分	14日分	5日分	14日分
国際	60.3 (12.0)	55.8 (11.5)	55.2 (11.4)	58.3 (11.8)	58.2 (11.8)
経済	69.5 (12.8)	65.0 (12.3)	63.7 (12.1)	67.0 (12.5)	66.4 (12.4)
家庭	92.7 (16.8)	84.7 (15.8)	82.9 (15.6)	86.5 (16.1)	85.5 (15.9)
スポーツ	83.0 (15.2)	69.4 (13.6)	64.6 (13.0)	76.7 (14.5)	74.4 (14.2)
社会	64.5 (14.0)	60.2 (13.4)	58.3 (13.1)	63.0 (13.8)	61.6 (13.6)

実験手順を以下に示す．

Step.1　形態素数 5000,20000 の基本語彙を構築する．

Step.2　基本語彙でテストデータのカバー率を求める．

Step.3　基本語彙に数日間～数週間の適応化サンプル中に出現した固有名詞を高出現頻度順に追加し，固有名詞のカバー率を求める．

実験は，追加登録する形態素を5000に限定した場合と出現したすべてを登録する場合を行なった．実験結果を表6,7に示す．表中の括弧内の数値は出現した固有名詞をすべて登録した場合の数を示している．この結果より次のことが言える．

- 6ヶ月前の記事より直前の記事に出現する固有名詞を追加する方がカバー率が高い．これより，新しく出現した固有名詞の多くは直前の数日間に渡って出現していることが分かる．
- 追加する固有名詞の数を制限しない場合は，適応化サンプルが多いほどカバー率が高くなるのは当然だが，固有名詞の数を制限した場合でも，10日間より30日間の適応化サンプルを用いた方が，カバー率は少し高くなる．
- テストデータ全体でのカバー率を見て分かるように，固有名詞を追加することによるカバー率の上昇は高々2%程度である．このことは，基本語彙に登録されなかった単語(未知語)において，固有名詞の占める割合が低いことを示している(5000語彙に対しては約20%，20000語彙に対しては約25%)．

なお，固有名詞に限定せずに，出現頻度の多い形態素を登録した場合の結果を表8に示す．

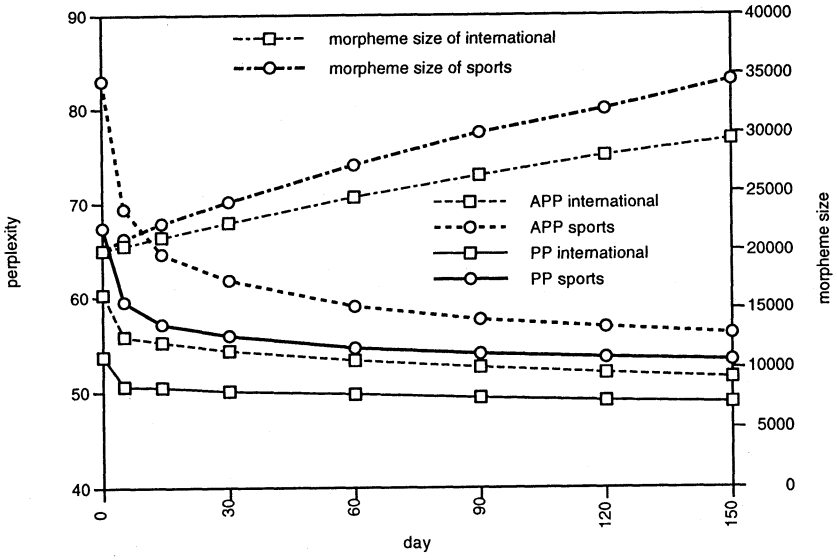

図 2　MAP推定の日数とパープレキシティの関係 (20000形態素)

表 6　固有名詞の追加登録によるテストデータ全体でのカバー率[%]の変化

基本語彙	追加語彙		直前の記事で適応			6カ月前の記事で適応		
サイズ	サイズ	適応前	10日分	20日分	30日分	10日分	20日分	30日分
5000	5000	85.2	86.6	86.8	86.8	86.4	86.5	86.6
	制限なし	85.2	86.8	87.1	87.2	86.6	86.9	87.0
	(追加語彙)	—	(6860)	(11071)	(14380)	(7096)	(11353)	(14677)
20000	5000	95.2	95.7	95.7	95.8	95.5	95.5	95.6
	制限なし	95.2	95.7	95.9	96.0	95.5	95.7	95.8
	(追加語彙)	—	(5222)	(9166)	(12383)	(5403)	(9411)	(12655)

表 7　固有名詞の追加登録によるテストデータ中の固有名詞についてのカバー率[%]の変化

基本語彙	追加語彙		直前の記事で適応			6カ月前の記事で適応		
サイズ	サイズ	適応前	10日分	20日分	30日分	10日分	20日分	30日分
5000	5000	41.9	75.0	78.3	79.2	69.0	72.1	74.1
	制限なし	41.9	79.5	85.0	87.7	73.6	80.5	84.2
	(追加語彙)	—	(6860)	(11071)	(14380)	(7096)	(11353)	(14677)
20000	5000	69.9	81.3	81.8	82.9	76.7	77.8	78.5
	制限なし	69.9	81.8	85.7	87.9	77.5	81.9	84.9
	(追加語彙)	—	(5222)	(9166)	(12383)	(5403)	(9411)	(12655)

表8より，登録する単語を固有名詞に限定しない方がカバー率は大きくなることかがわかる．しかし，このような新しい登録単語のbigram,trigramの算出は困難なので固有名詞に限定した方が扱いやすいと思われる．また表8より，カバー率を98%にするためには直前に出現した形態素を中心とした55000形態素程度が必要なことがわかる．但し，面種別で学習すれば，20000〜30000形態素でも十分である(表 2,3参照)．

表 8　高出現頻度の形態素の追加登録によるテストデータ全体でのカバー率[%]の変化(品詞の限定なし)

基本語彙サイズ	追加語彙サイズ	適応前	直前の記事で適応			6カ月前の記事で適応		
			10日分	20日分	30日分	10日分	20日分	30日分
5000	5000	85.2	90.8	91.0	91.1	90.0	90.3	90.5
	制限なし(追加語彙)	85.2 —	96.8 (34012)	97.9 (49539)	98.4 (60665)	96.2 (34287)	97.5 (49457)	98.1 (60314)
20000	5000	95.2	96.1	96.2	96.2	95.8	95.9	95.9
	制限なし(追加語彙)	95.2 —	97.3 (20906)	98.0 (35158)	98.4 (45966)	96.9 (21014)	97.7 (34999)	98.1 (45559)

4　定型表現

新聞テキスト文には，定型表現が多いことに着目し，これらの高頻出定型表現を1形態素として捉えた上で，言語モデルを構築すれば，より精度の良いモデルが出来ると考られる．

今回の実験では，定型表現を抽出するアルゴリズムとして，池原らの提案した方法(池原, 白井, 河岡 1995)を用いる．エントロピー基準で連語を抽出する方法も考えられるが(Giachin 1995; 政瀧 1995; 森, 山地, 長尾 1997b)，今回は簡略化のため出現頻度に着目した．どのような基準で連語を抽出し，言語モデルを構築するかは興味ある課題であるが，手法による実質的な差は少ないと思われる(中川 1998)．この方法では，最長一致の文字列抽出(ある文字列が抽出されたとき，その文字列に含まれる部分文字列は統計量を求める際にはこの部分文字列を定型表現とはカウントしない)を条件とし，任意の長さ以上，任意の使用頻度以上の表現を，もれなく自動的に抽出する．文献(池原 1995)では文字列単位で抽出していたが，これを形態素単位で抽出するように変更した．抽出例を表 9に示す．

表 9　定型表現抽出例

連語数	定型表現(頻度)
2	て/いる (318691) は/ない (56333) 東京/都 (23452) 大統領/は (14647) 国民/の (9909)
3	し/て/いる (106121) に/よる/と (24093) に/なっ/て (19718) 話し/て/いる (6130) 記者/会見/し (4297)

4.1 標準言語モデル

標準言語モデルは，表 1に示した全面種の学習用データから作成した表 2のモデルを用いる．まず，RWC(井佐原 *et al.* 1995)の毎日新聞形態素解析結果を用いて，出現頻度が上位20000番目までの形態素を語彙として辞書に登録した．言語モデルの構築には，CMU SLM Toolkit Ver.1を用いた．バックオフ・スムージングにはGood-Turing推定を用いた．

4.2 定型表現を用いた言語モデル

定型表現を用いた言語モデル構築のための手順を以下に示す．

Step.1 定型表現抽出

RWCの毎日新聞形態素解析結果に対して，定型表現抽出プログラムを実行し，連結数2または3の定型表現を抽出する．

Step.2 頻度の計算

定型表現を用いる前のトレーニングデータから，各形態素の頻度リストを求める．上位15000番目くらいの形態素の出現頻度が50回であるので，定型表現の出現頻度が50回以上のものを新しい形態素候補として用いることにする．

Step.3 定型表現の連結

Step.2の定型表現を用い，トレーニングデータ内の定型表現を図 3のように1つの単語にまとめる．

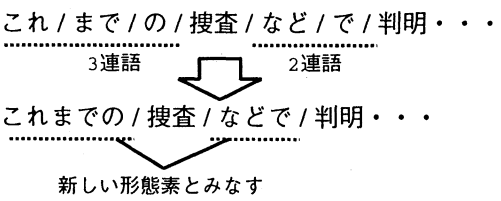

図 3　形態素の連結例

Step.4 語彙サイズ20000の辞書作成(1回目)

トレーニングデータから出現頻度の多い順に20000を求め，語彙サイズ20000の辞書を作成する．このとき，上位20000の辞書に登録された定型表現は2連結で9430個，3連結で9357個

(このうち2連語が7010，3連語が2347)である．登録されなかった定型表現が多数あるので，これは未知語の数を増やすだけなので図4のようにもとの形態素に分解しておく．

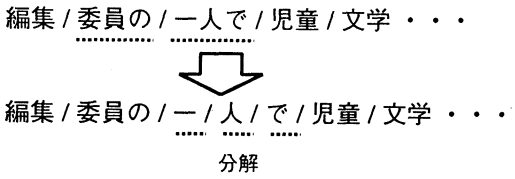

図 4　形態素の分解例

Step.5 語彙サイズ20000の辞書作成(2回目)

分解後のトレーニングデータから，もう一度語彙サイズ20000の辞書を作成する．これは，Step.3で定型表現を分解したことによって形態素の出現頻度が変わってしまうためである．当然ここでも登録されない定型表現がでてくる．登録された定型表現は2連結で8944個，3連結で9282個(このうち2連語が6967，3連語が2315)になった．ここでも，登録されなかった定型表現はもとの形態素に分解する．

厳密に行なうなら，辞書作成と定型表現の分解といった作業を繰り返し行ない，辞書に登録される定型表現の数が収束するまで行わないといけないが，今回は1回だけしか行なっていない．

Step.6 言語モデルの構築

CMU SLM Toolkitを用いてトレーニングデータから，語彙サイズ20000の辞書を作成し，bigram,trigram言語モデルを構築する．

4.3　評価実験

評価を行なう時，注意しなければならないことは，いずれの比較対象に対しても同じ定義の1形態素あたりのパープレキシティを求めないといけないということである．通常パープレキシティを求める式は**bigram**の場合で，

$$PP_0 = \sqrt[M]{\prod_{i=1}^{M} P(w_i|w_{i-1})^{-1}} \tag{8}$$

であるが，これは1連結形態素(定型表現として形態素を連結したもの)あたりのパープレキシティを求めている．形態素を連結する前の従来の1形態素あたりのパープレキシティを求める

には,

$$PP_1 = \sqrt[N]{\prod_{i=1}^{M} P(w_i|w_{i-1})^{-1}} \tag{9}$$

を用いなければならない．ここで

M:定型表現を1つの形態素としたときの連結形態素と従来の形態素の総数

N:定型表現を使わなかったときの従来の形態素の総数

また，定型表現は述語表現に多く現れるため，それらの形態素は比較的短いものが多い．そのため形態素単位のパープレキシティでは全体に及ぼす影響が大きいと考えられる．そこで同時に文字単位のパープレキシティも求めた．

標準言語モデルと，前節に述べた方法で定型表現を用いた言語モデルを構築し，その評価を行なった．トレーニングデータには，標準言語モデル作成の場合と同じ，表1の学習用データを用いている．テストデータには，標準言語モデル，定型表現を用いた言語モデルともに表1の評価用データを使用した．

実験結果を表10と図5に示す．まず，bigramモデルでは，トレーニングデータに関しては約半分，テストデータに関しては約3割，パープレキシティが減少しているのがわかる．しかし，trigramモデルではトレーニングデータでは効果があったが，テストデータに対しては大きな効果が得られなかった．これは，語彙サイズを一定にしたため，定型表現を登録したためにもとの語彙から省かれた単語が未知語となったのが原因であると考えられる．実際，定型表現を用いた場合，定型表現を用いなかった場合と比べて，未知語の種類数が約8000個増加している．

次に，標準言語モデルを作成した時の語彙サイズ20000の辞書に，2および3連結の定型表現をそれぞれ高出現頻度順で上位2000個,5000個分を追加した場合の辞書で言語モデルを構築した．その評価結果を表11,12に示す．ここで表11,12の“定型表現の連結なし”は通常の形態素を22000個および25000個用いた時の結果である．この言語モデルの作成方法の場合でも，パープレキシティの改善が見られた．bigramでは定型表現を用いることにより，補正パープレキシティも大幅に減少している．また，定型表現5000個追加のものの方が，定型表現2000個追加のものと比べて，語彙サイズが大きいのにも関わらず，パープレキシティが減少している．これより，出来るだけ多くの定型表現を辞書に登録すれば良いということが言える．

以上より，trigramではトレーニングデータに対しては大きな効果があったが，テストデータに対しては効果がなかった．これはトレーニングデータの不足によるものと考えられる．一方，bigramでは大きな効果があった．同じパラメータ数(bigram)でもパープレキシティが小さいモデルが構築できたことは，これを大語彙連続音声認識の第1パスに使用すると認識率の向上に繋がると考えられる．

表 10　定型表現の評価結果(語彙サイズ20000)
(括弧内は文字単位のパープレキシティ)

データセット		トレーニングデータ			テストデータ		
定型表現		なし	2連結	3連結	なし	2連結	3連結
bigram	PP	91.0 (16.3)	57.5 (12.2)	52.2 (11.5)	105.5 (17.9)	75.6 (14.5)	73.3 (14.3)
	APP	136.7 (20.9)	121.3 (19.4)	113.1 (18.6)	156.0 (22.8)	160.8 (23.2)	151.6 (22.4)
trigram	PP	29.7 (8.1)	20.1 (6.4)	13.2 (4.9)	61.3 (12.8)	65.1 (13.3)	55.4 (12.0)
	APP	44.6 (10.5)	42.3 (10.1)	28.6 (7.9)	90.7 (16.3)	131.5 (20.5)	114.6 (18.8)

表 11　定型表現の評価結果(語彙サイズ20000+2000)
(括弧内は文字単位のパープレキシティ)

データセット		トレーニングデータ			テストデータ		
定型表現		なし	2連結	3連結	なし	2連結	3連結
bigram	PP	92.7 (16.4)	73.7 (14.3)	75.9 (14.5)	108.5 (18.2)	93.9 (16.6)	98.2 (17.1)
	APP	133.4 (20.6)	110.7 (18.3)	113.9 (18.7)	153.7 (22.6)	138.9 (21.2)	145.3 (21.8)
trigram	PP	29.7 (8.1)	19.5 (6.3)	18.8 (6.1)	62.8 (13.0)	62.8 (13.0)	64.6 (13.2)
	APP	42.7 (10.2)	29.3 (8.1)	28.2 (7.8)	89.0 (16.1)	91.9 (16.4)	95.5 (16.8)

表 12　定型表現の評価結果(語彙サイズ20000+5000)
(括弧内は文字単位のパープレキシティ)

データセット		トレーニングデータ			テストデータ		
定型表現		なし	2連結	3連結	なし	2連結	3連結
bigram	PP	94.8 (16.7)	66.0 (13.3)	66.5 (13.4)	111.9 (18.5)	89.6 (16.2)	93.6 (16.6)
	APP	129.6 (20.2)	99.1 (17.1)	99.9 (17.2)	151.4 (22.4)	132.6 (20.6)	138.5 (21.2)
trigram	PP	30.2 (8.2)	16.6 (5.7)	14.9 (5.3)	64.6 (13.2)	63.2 (13.0)	65.9 (13.4)
	APP	40.5 (9.9)	24.9 (7.3)	22.4 (6.8)	87.5 (15.9)	93.5 (16.6)	97.6 (17.0)

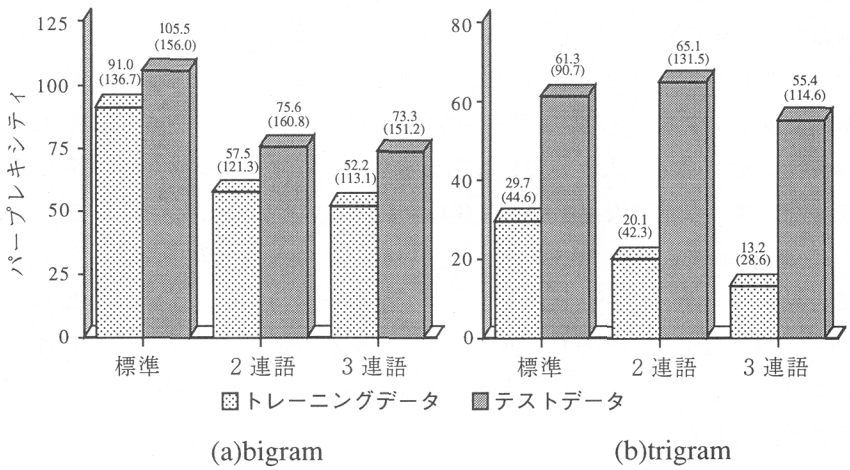

図 5　定型表現の評価結果 [注:() 内の数値は補正パープレキシティを示す]

5　まとめ

本研究では，毎日新聞記事データベースを用いた過去の記事による言語モデルのタスクへの適応化と抽出した定型表現を用い，N-gram言語モデルを構築する方法を検討した．

まず言語モデルのタスクへの適応化については，実験の結果，6カ月前の数日間の記事より直前の数日間の記事で適応化した方がパープレキシティが小さくなった．このことは言語モデルがジャンルだけでなく時間にも依存するものであることを示すものである．ただ，適応化サンプルの量を多くするほどパープレキシティが小さくなる傾向があり，N-gramベースでの言語モデルを少量サンプルで適応化させることは限界があると考えられる．

次に定型表現を抽出し，これを用いたN-gram言語モデルを構築した．定型表現を用いた言語モデルを作成することで，bigramモデルに関しては，テストデータに対し約3割程度パープレキシティを低く押えるとこができ，言語モデルの有効性を示すことができた．しかし，trigramではトレーニングデータの量が不十分だったため，トレーニングデータでは効果があったがテストデータに対しては効果が得られなかった．トレーニングデータの量をもっと増やし，本方法の有効性を調べる必要がある．また，本研究では言語モデルの有効性をパープレキシティで評価したが，実際の音声認識で確認する必要がある(赤松, 甲斐, 中川 1998)．

なお，NHKのニュース原稿に対する経時変化の適応化や定型表現の導入による言語モデルに関しては文献(小林, 今井, 安藤 1997; 西崎, 中川 1998)を参照されたい．

参考文献

赤松裕隆, 中川聖一 (1997). “新聞記事のトライグラムによるモデル化と適応化” 言語処理学会, 第3回年次大会 D5-2,533-536.

赤松裕隆, 甲斐充彦, 中川聖一 (1998). “新聞・ニュース文の大語彙連続音声認識” 情報処理学会, 音声言語情報処理 SLP-21-11, 97–104.

Federico, M.(1997). “Baysian Estimation Methods for N-gram Language Model Adaptation.” Proc.ICSLP-96, 240–243.

Giachin, E.P.(1995). “Phrase bigrams for continuous speech recognition.” Proc.ICASSP, 225–228.

池原悟, 白井諭, 河岡司 (1995). “大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法.” 情報処理学会論文誌, Vol.36, No.11, 2584–2596.

井佐原均, 元吉文男, 徳永健伸, 橋本三奈子, 荻野紫穂, 豊浦潤, 岡隆一 (1995). “RWCにおける品詞情報付きテキストデータベースの作成” 言語処理学会, 第1回年次大会 B3-1, 181 – 184.

伊藤彰則, 好田正紀 (1996). “対話音声認識のための事前タスク適応の検討.” 情報処理学会, 音声言語情報処理 SLP-14-13.

伊藤克亘 *et al.*,(1997). “大語彙日本語連続音声認識研究基盤の整備—学習・評価テキストコーパスの作成—.” 情報処理学会, 音声言語情報処理 SLP-18-2,7–12.

伊藤克亘, 松岡達雄, 竹沢寿幸, 武田一哉, 鹿野清宏 (1996). “大語彙連続音声認識研究のためのテキストデータ処理.” 日本音響学会秋季講演論文集 3-3-10, 105–106.

甲斐充彦, 伊藤敏彦, 山本一公, 中川聖一 (1997). “自然な発話を対象としたパソコン／ワークステーション用連続音声認識ソフトウェア.” 日本音響学会秋季講演論文集 2-Q-30, 175–176.

北研二, 小倉健太郎, 森元逞, 矢野米雄 (1995). “仕事量基準を用いたコーパスからの定型表現の自動抽出.” 情報処理学会論文誌, Vol.34, No.9, 1937–1943.

小林彰夫, 今井亨, 安藤彰男 (1997). “ニュース音声認識用言語モデルの学習期間の検討.” 電子情報通信学会, 音声技報 SP97-48, 29–36.

小林紀彦, 中野裕一郎, 和田陽介, 小林哲則 (1998). “統計的言語モデルにおける高頻度形態素連鎖の辞書登録に関する一考察.” 情報処理学会, 音声言語情報処理 SLP-20-5, 33–38.

Kuhn, R., Mori, R.(1990). “A cache-based natural language model for speech recognition.” IEEE Trans Pattern Analysis and Machine Intelligence, Vol.12, No.6, 570–583.

Marlin, S.C., Liermann, J.,(1997). “Adaptive topic dependent language modelling using word-based varigrams.” Proc.EuroSpeech, 1447–1450.

政瀧浩和, 松永昭一, 匂坂芳典 (1995). “連続音声認識のための可変長連鎖統計言語モデル.” 電子情報通信学会, 音声技報 SP95-73, 1–6.

政瀧浩和, 匂坂芳典, 久木和也, 河原達也 (1997). “MAP推定を用いたN-gram言語モデルのタスク適応.” 電子情報通信学会, 音声技報 SP96–103, 59–64.

松永昭一, 山田智一, 鹿野清宏 (1991). “音節連鎖統計情報のタスク適応化.” 情報処理学会第42回全国大会 (2) 6D-5, 114-115.

森信介, 山地治 (1997). “日本語情報量の上限の推定.” 情報処理学会論文誌, Vol.38, No.11, 2191–2199.

森信介, 山地治, 長尾眞 (1997). “予測単位の変更によるn-gramモデルの改善.” 情報処理学会, 音声言語情報処理 SLP-19-14, 87–94.

中川聖一 (1992). “情報理論の基礎と応用.” 近代科学社.

中川聖一 (1998). “音声認識のための統計的言語モデル.” 日本音響学会春季講演論文集 1-6-11, 23–26.

中川聖一, 赤松裕隆 (1998). “未知語を含む文集合のパープレキシティの算出法ー新補正パープレキシティ.” 日本音響学会秋季講演論文集 2-1-13, 63–64.

西村雅史, 伊藤伸泰, 山崎一孝, 萩野紫穂 (1998). “単語を認識単位とした日本語の大語彙連続音声認識.” 情報処理学会, 音声言語情報処理 SLP-20-3,17–24.

西崎博光, 中川聖一 (1998). “音声認識のための定型表現を用いた言語モデルの検討.” 言語処理

学会, 第4回年次大会 C4-3, 520-523.

小黒玲, 高木一幸, 橋本顕示, 尾関和彦 (1998). “ニュース音声認識のための言語モデルの比較.” 日本音響学会春季講演論文集 1-6-22, 47–48.

大附克年, 吉田航太郎, 松岡達雄, 古井貞照 (1997). “高次n-gramを用いた大語彙連続音声認識の検討.” 日本音響学会春季講演論文集 2-6-2, 47–48.

大附克年, 森岳至, 松岡達雄, 古井貞照, 白井克彦 (1995). “新聞記事を用いた大語彙連続音声認識の検討.” 電子情報通信学会, 音声技報 SP95-90, 63–68.

Rosenfeld, R.(1995). “The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation.” Proc. ARPA Spoken Language Systems Technology Workshop, 47–50.

新納浩幸, 井佐原均 (1995). “疑似Nグラムを用いた助詞的定型表現の自動抽出.” 情報処理学会論文誌, Vol.36, No.1, 32–40.

Ueberla, J.(1994). “Analysing a simple language model - some general conclusion for language models for speech recognition.” Computer Speech and Language, Vol.8, No.2, 153–176.

Woodland, P.C., Cales, M.J.F., Pye, D., Young, S.J.(1997). “The development of the 1996 HTK broadcast news transcription systems.” Proc.Speech Recognition Workshop, 73–78.

略歴

中川聖一: 1976年京都大学大学院博士課程修了. 同年京都大学情報助手. 1980年豊橋技術科学大学情報工学系講師. 1983年助教授. 1990年教授. 1985~86年カーネギメロン大学客員研究員. 音声情報処理, 自然言語処理, 人工知能の研究に従事. 工博. 1977年電子情報通信学会論文賞, 1988年度IETE最優秀論文賞受賞.

赤松 裕隆: 1997年豊橋技術科学大学情報工学課程卒業. 現在, 同大学研究科修士課程情報工学専攻在学中. 言語モデルに関する研究に従事.

西崎 博光: 1998年豊橋技術科学大学情報工学課程卒業. 現在, 同大学研究科修士課程情報工学専攻在学中. 言語モデルに関する研究に従事.

(1998 年 4 月 3 日 受付)

(1998 年 7 月 27 日 採録)